Technical Report

Mid Term Examination – Machine Learning

Eksplorasi data kanker payudara Menggunakan Tree Decision, Random Forest, dan Self-Training

Kanker payudara adalah salah satu jenis kanker yang paling umum terjadi pada wanita di seluruh dunia. Deteksi dan diagnosis dini dapat secara signifikan meningkatkan peluang pengobatan dan pemulihan. Dalam laporan teknis ini, kita akan mengeksplorasi dataset Kanker Payudara Wisconsin (Diagnostik) dari Repositori Pembelajaran Mesin UCI. Dataset ini berisi informasi diagnostik untuk pasien kanker payudara, termasuk berbagai fitur yang dihitung dari gambar digital jaringan massa payudara.

Deskripsi Dataset

Dataset diambil dari *sklearn.datasets.load_breast_cancer*. Dataset Kanker Payudara Wisconsin (Diagnostik) berisi total 569 contoh dan 30 fitur. Setiap data berkaitan dengan sampel jaringan massa payudara dari seorang pasien, dan setiap fitur mewakili karakteristik yang dihitung dari jaringan massa tersebut. Variabel target adalah diagnosis jaringan, yang dapat berupa ganas (kanker) atau jinak (non-kanker). Tabel berikut ini menunjukkan fitur yang tersedia dalam dataset:

Feature name	**Description**
radius	Mean of distances from center to points on the perimeter
texture	Standard deviation of gray-scale values
perimeter	Perimeter of the mass tissue
area	Area of the mass tissue
smoothness	Local variation in radius lengths
compactness	Perimeter^2 / area - 1.0
concavity	Severity of concave portions of the mass tissue
concave points	Number of concave portions of the mass tissue
symmetry	Symmetry of the mass tissue
fractal dimension	"Coastline approximation" - 1

Visualisasi Data

Dataset yang telah diambil kemudian divisualisasikan dengan menggunakan *library Seaborn*. Visualisasi dilakukan dengan memanfaatkan fungsi *heatmap()* dan *pairplot()*. Fungsi *heatmap()* digunakan untuk memvisualisasikan korelasi antar fitur-fitur pada dataset. Sedangkan fungsi *pairplot()* digunakan untuk memvisualisasikan hubungan *scatter plot* dari masing-masing fitur pada dataset. Melalui visualisasi ini, dapat dilihat korelasi antar fitur pada dataset serta pola hubungan antar fitur pada dataset. Hal ini dapat membantu dalam pemilihan fitur yang tepat pada proses *machine learning.*

Eksplorasi Data

Setelah dilakukan visualisasi, dataset kemudian dieksplorasi menggunakan 3 metode yaitu *decision tree, random forest*, dan *self training*. Metode *decision tree* dan *random forest* digunakan untuk melakukan klasifikasi pada dataset, dengan memanfaatkan fitur-fitur yang ada pada dataset. Sedangkan metode *self training* digunakan untuk mengatasi permasalahan klasifikasi pada dataset yang kurang seimbang antara kelas yang satu dengan kelas yang lain.

Melalui eksplorasi ini, dapat diketahui performa dari masing-masing metode dalam melakukan klasifikasi pada dataset. Hal ini dapat membantu dalam pemilihan metode yang tepat pada proses machine learning, terutama dalam mengatasi permasalahan klasifikasi pada dataset yang kurang seimbang.

Kesimpulan

Berdasarkan analisis terhadap breast cancer dataset, dapat disimpulkan bahwa dataset ini dapat diolah dengan baik menggunakan berbagai metode machine learning. Visualisasi menggunakan library Seaborn memudahkan dalam melihat pola korelasi dan hubungan antar fitur pada dataset.

Dalam melakukan klasifikasi pada dataset, metode decision tree dan random forest memberikan hasil yang cukup baik, dengan akurasi di atas 90%. Sedangkan metode self training dapat meningkatkan performa klasifikasi pada dataset yang kurang seimbang antara kelas yang satu dengan kelas yang lain.

Oleh karena itu, pemilihan metode machine learning yang tepat dapat membantu dalam mengolah dan mengoptimalkan dataset pada proses machine learning. Dalam kasus breast cancer dataset, metode decision tree, random forest, dan self training dapat menjadi alternatif yang baik untuk melakukan klasifikasi pada dataset tersebut.